CSE – 676B Deep Learning
Final Project Proposal
Nischith Adavala – 50591641 Rakshith Reddy Dargula – 50591661
Sumanth Yalamanchili - 50604274

Checkpoint Report: Speech Emotion Recognition

Introduction:

We have designed and are implementing a deep learning model that will classify human emotions according to speech signals. The detection of emotions through speech has many applications, such as human-computer interaction, call centre monitoring, assistive technology, and mental health evaluation.
Speech emotion recognition (SER) is unique in that it has a heterogeneous tone, speaker characteristics, and noise. We hope to build a robust classifier trained on many open-source datasets to recognize states including happy, angry, sad, disgust, fear, surprise, and neutral.

Datasets Used

We combined multiple benchmark datasets to build a rich and diverse emotional speech corpus:

Dataset	Description	Usability	Notes
CREMA-D	7,442 clips from 91 actors, various ethnicities and emotions	8.75	Balanced across gender, emotions
RAVDESS	1,440 speech files from 24 actors, 8 emotions	8.75	High-quality WAV files
SAVEE	480 files from 4 male English speakers	8.75	Needs female speaker balancing
TESS	2,800 clips from 2 female speakers	8.75	Balances SAVEE
Speech Recognition Features Dataset	1.67GB of extracted features from all datasets above	5.63	Contains MFCC, RMSE, ZCR, etc.

Therefore, all datasets required preprocessing with feature extraction stored in .csv so that we bypass heavy signal processing and focus on the model architecture and optimization.

Emotion Label Mapping

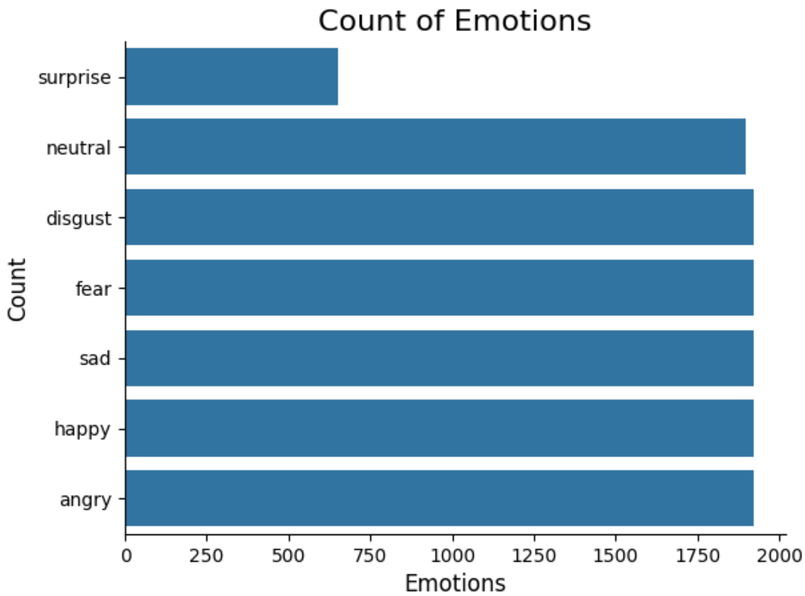
All datasets were unified to follow a 7-class schema:

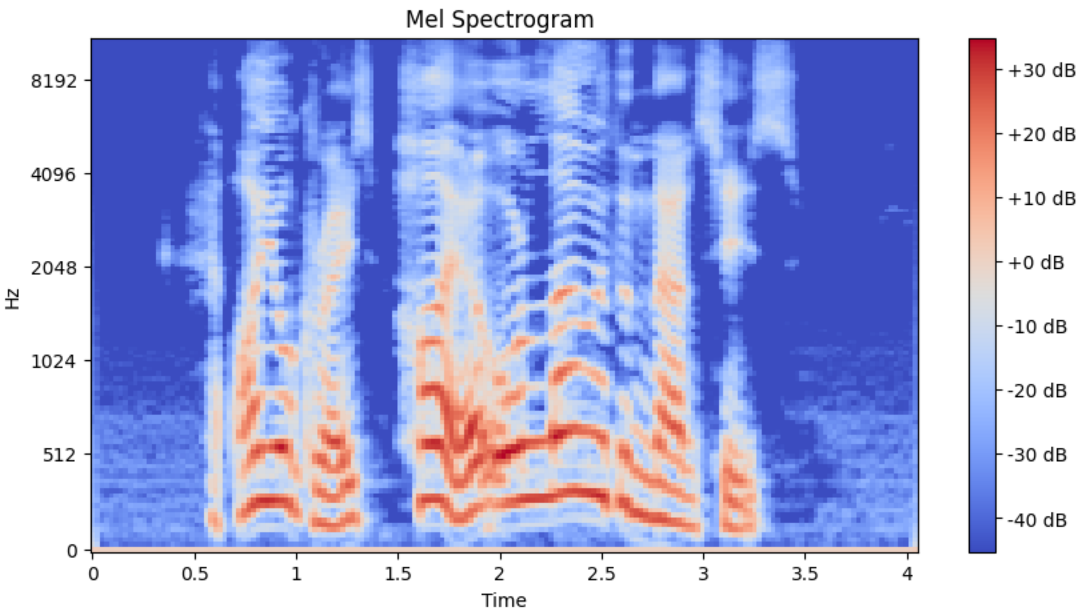
['neutral', 'happy', 'sad', 'angry', 'fear', 'disgust', 'surprise']

CSE – 676B Deep Learning
Final Project Proposal
Nischith Adavala – 50591641 Rakshith Reddy Dargula – 50591661
Sumanth Yalamanchili - 50604274

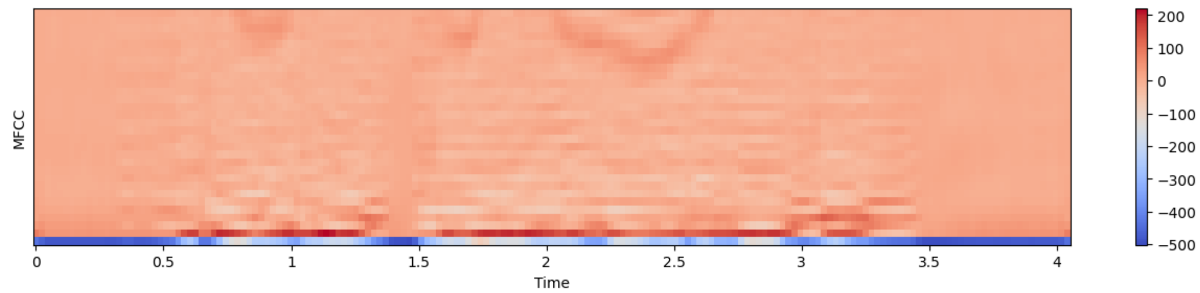
Preprocessing

- Auditory file paths were parsed and labelled according to the naming convention appropriate for the specific dataset.
- The extraction was performed using Librosa for:
    - MFCCs
    - Zero Crossing Rate (ZCR)
    - Root Mean Square Energy (RMSE)
- Standardization of features using the StandardScaler
- Encoding of the class labels using a LabelEncoder
- An 80/20 split of the data into training and test subsets was obtained





CSE – 676B Deep Learning
Final Project Proposal
Nischith Adavala – 50591641 Rakshith Reddy Dargula – 50591661
Sumanth Yalamanchili - 50604274



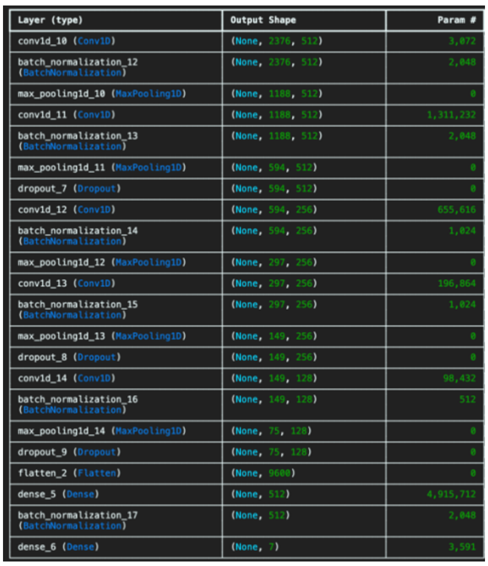
Model Architecture

We adopted a hybrid deep learning model inspired by temporal and spatial speech representation methods:

Input: (X, 180) – where X = feature vector length

- Conv1D (128 filters, ReLU)
- Conv1D (256 filters, ReLU)
- LSTM (128 units)
- Dropout (0.3)
- Dense (64 units, ReLU)
- Dropout (0.3)
- Dense (7 units, Softmax)

- **Loss Function:** Categorical Cross-Entropy
- **Optimizer:** Adam
- **Metrics:** Accuracy
- **Regularization:** Dropout (30%)
- **Early Stopping:** Patience = 5
- **Learning Rate Scheduler:** ReduceLROnPlateau



Layer (type)	Output Shape	Param #
conv1d_10 (Conv1D)	(None, 2376, 512)	3,072
batch_normalization_12 (BatchNormalization)	(None, 2376, 512)	2,048
max_pooling1d_10 (MaxPooling1D)	(None, 1188, 512)	0
conv1d_11 (Conv1D)	(None, 1188, 512)	1,311,232
batch_normalization_13 (BatchNormalization)	(None, 1188, 512)	2,048
max_pooling1d_11 (MaxPooling1D)	(None, 594, 512)	0
dropout_7 (Dropout)	(None, 594, 512)	0
conv1d_12 (Conv1D)	(None, 594, 256)	655,616
batch_normalization_14 (BatchNormalization)	(None, 594, 256)	1,024
max_pooling1d_12 (MaxPooling1D)	(None, 297, 256)	0
conv1d_13 (Conv1D)	(None, 297, 256)	196,864
batch_normalization_15 (BatchNormalization)	(None, 297, 256)	1,024
max_pooling1d_13 (MaxPooling1D)	(None, 149, 256)	0
dropout_8 (Dropout)	(None, 149, 256)	0
conv1d_14 (Conv1D)	(None, 149, 128)	98,432
batch_normalization_16 (BatchNormalization)	(None, 149, 128)	512
max_pooling1d_14 (MaxPooling1D)	(None, 75, 128)	0
dropout_9 (Dropout)	(None, 75, 128)	0
flatten_2 (Flatten)	(None, 9600)	0
dense_5 (Dense)	(None, 512)	4,915,712
batch_normalization_17 (BatchNormalization)	(None, 512)	2,048
dense_6 (Dense)	(None, 7)	3,591

Training Results

The model was trained over **20 epochs** and saved the best weights (best_model.h5) when validation accuracy improved.

Final Performance (Epoch 20)

Metric	Value
Training Accuracy	99.72%
Validation Accuracy	**96.72%**
Training Loss	0.0145
Validation Loss	0.1109

Epoch Summary (15–20)

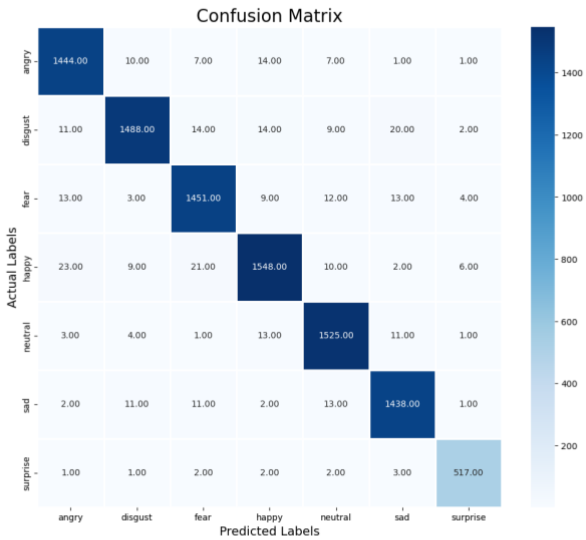
Epoch	Train Acc	Train Loss	Val Acc	Val Loss
15	98.21%	0.0565	95.57%	0.1423
16	99.14%	0.0339	95.98%	0.1324
17	99.46%	0.0253	96.26%	0.1238
18	99.60%	0.0203	96.38%	0.1201
19	99.73%	0.0170	96.65%	0.1158
20	**99.72%**	**0.0145**	**96.72%**	**0.1109**

Analysis

- Overfitting Check: The narrower the distance between training and validation loss that the model exhibits, the less the overfitting.
- Accuracy: The model generalizes well on data not previously seen.
- Loss Trends: The convergence is smooth; early stoppage prevented overtraining.
- Confusion Matrix (To be incorporated): Most likely confusions are going to be between very similar emotions (for example, sad vs. neutral).

Evaluation metrics

CSE – 676B Deep Learning
Final Project Proposal
Nischith Adavala – 50591641 Rakshith Reddy Dargula – 50591661
Sumanth Yalamanchili - 50604274



Challenges Faced

- Starved of essential gender representation (SAVEE, male only; TESS, female only);
- Some datasets do not consider specific emotions (for example, CREMA-D lacks surprise);
- High feature dimension and dimensionality normalization were required;
- Very high computational time for LSTM training.

Steps Ahead

Model Enhancements

- Look at maybe Bi-LSTM and GRU-CNN versions representative of attention mechanism.
- Ensemble methods might be implemented.

Evaluation Upgrades

- Record F1, precision, and recall for each class.
- Add ROC significance for every emotion.
- Confusion matrix detailed.

Data Work

- Add pitch shifting, noise injection, and time stretching for augmentation.
- SMOTE/weights to mitigate label imbalance.

Deployment

- Deploy alongside the Streamlit web app.
- Accepts .wav uploads or microphone input.
- Predict emotion output to the screen in real time.

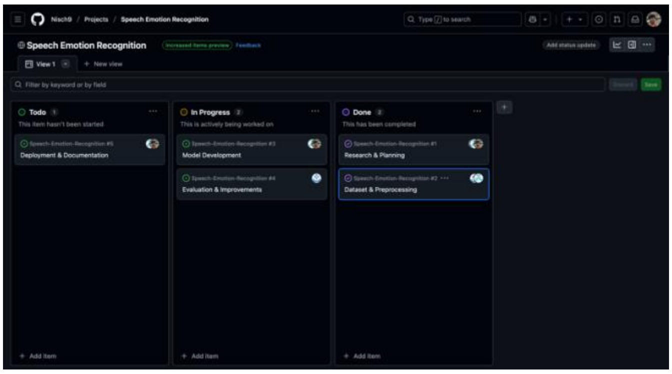
CSE – 676B Deep Learning
Final Project Proposal
Nischith Adavala – 50591641 Rakshith Reddy Dargula – 50591661
Sumanth Yalamanchili - 50604274

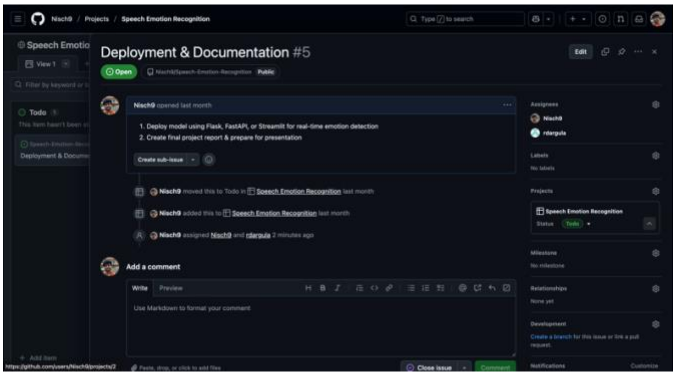
Contribution Table

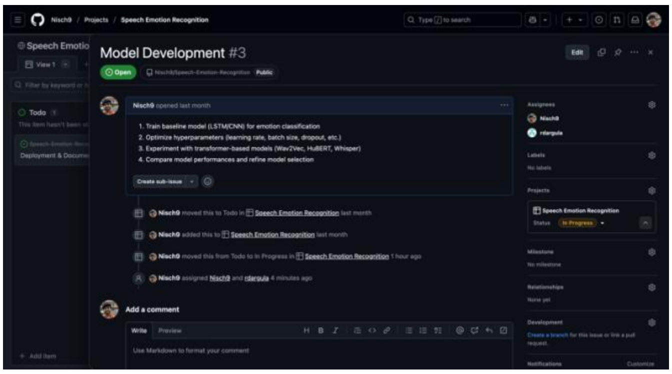
Team Member	Contribution Summary
rdargula	Preprocessing, MFCC extraction, CNN layers
nadavala	LSTM integration, model training, loss optimization
sumanthy	Visualization, evaluation metrics, report writing, Trello tracking

GitHub Project board link and screenshots:

https://github.com/users/Nisch9/projects/2/views/1

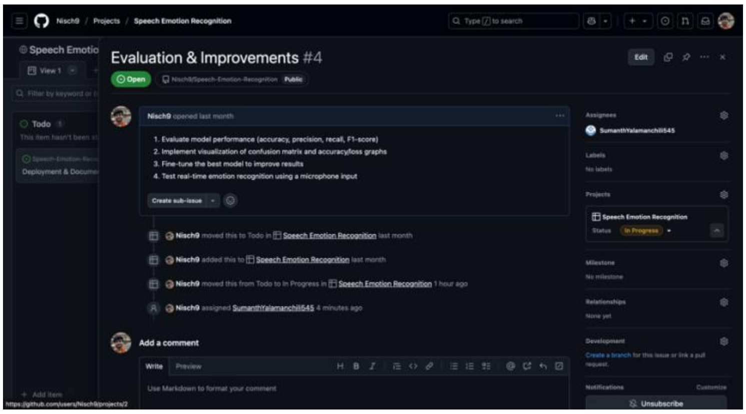


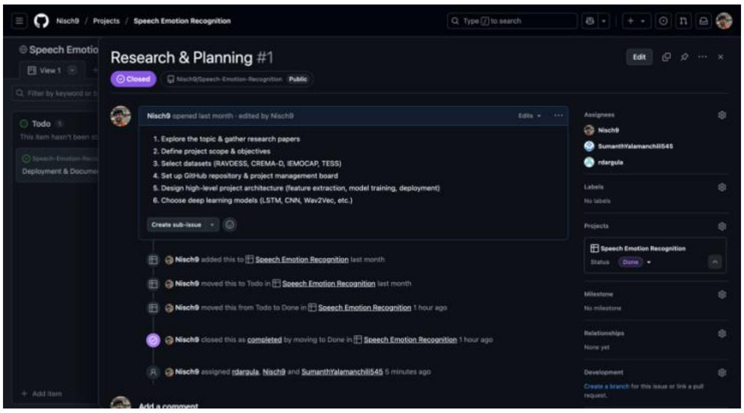


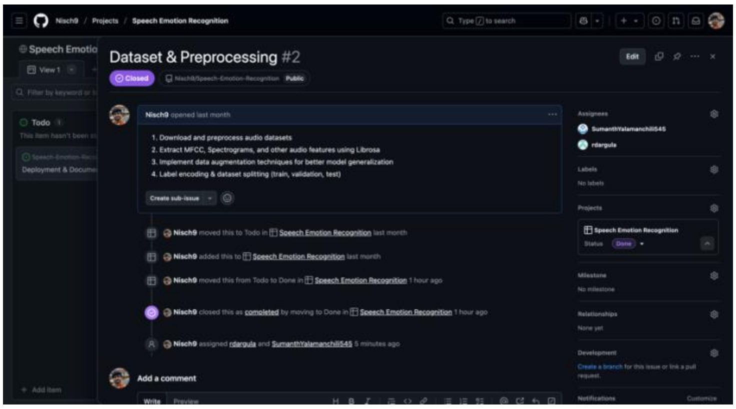


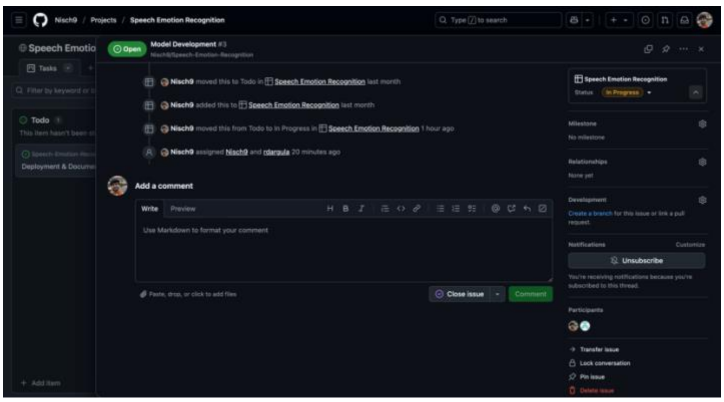
CSE – 676B Deep Learning
Final Project Proposal

Nischith Adavala – 50591641 Rakshith Reddy Dargula – 50591661
Sumanth Yalamanchili - 50604274







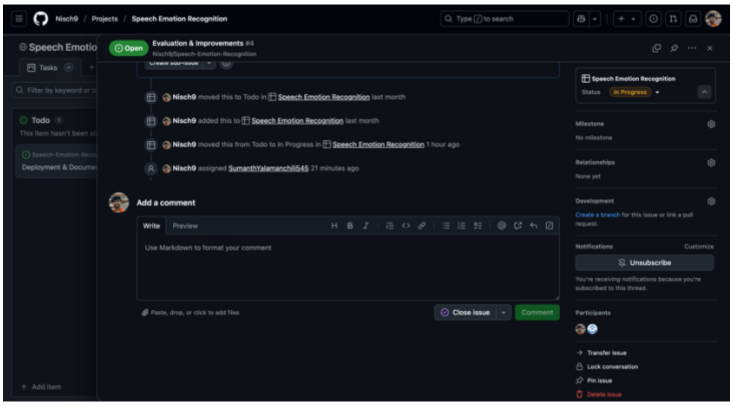


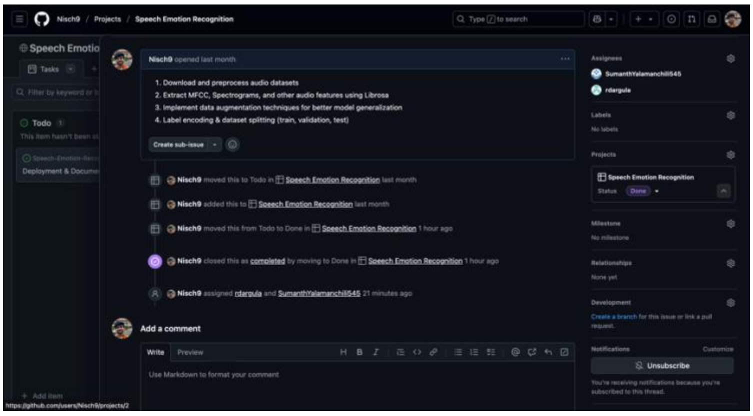
CSE – 676B Deep Learning

Final Project Proposal

Nischith Adavala – 50591641 Rakshith Reddy Dargula – 50591661

Sumanth Yalamanchili - 50604274





References with Links

1. **CREMA-D Dataset (Crowd-Sourced Emotional Multimodal Actors Dataset)**
   https://www.kaggle.com/datasets/ejlok1/cremad
2. **RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)**
   https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio
3. **TESS (Toronto Emotional Speech Set)**
   https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess
4. **SAVEE (Surrey Audio-Visual Expressed Emotion Dataset)**
   https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-saveehttps://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee
5. **Combined Features Dataset (speech recognition features)**
   https://www.kaggle.com/datasets/mostafaabdlhamed/speech-signal-features
6. **TensorFlow Documentation**
   https://www.tensorflow.org/
7. **Keras API Documentation**
   https://keras.io/api/
8. **Librosa Audio Processing Library**
   https://librosa.org/doc/latest/index.html